DOCUMENT RESUME

ED 194 536 TM 800 664

AUTHOR Waller, Michael I.
TITLE An Objective Procedure for Comparing the One, Two, and Three-Parameter Logistic Latent Trait Models.
PUB DATE Apr 80
NOTE 18p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Latent Trait Theory: *Mathematical Models: *Maximum Likelihood Statistics: *Measurement Techniques
IDENTIFIERS Monte Carlo Methods: Rasch Model: Two Parameter Model

ABSTRACT

An objective method based on the likelihood ratio procedure is presented for use in selecting a measurement model from among the RASCH, 2-parameter and 3-parameter logistic latent trait models. The procedure may be applied in a straightforward manner to aid in choosing between the 2-parameter and the Rasch models. When choosing between the 3- and 2-parameter models, one must use the values of the parameters which maximize the likelihood on a restricted range incorporating the guessing parameter and the number of alternatives. Also, given the inherent problem with the 3-parameter model of possible non-unique maximum likelihood parameter estimates and viable alternative procedures for handling random guessing, the recommendation is made that the 3-parameter model should be held in abeyance as a viable measurement model at this time. (Author/RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED194536

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

An Objective Procedure for Comparing the One, Two, and Three-Parameter Logistic Latent Trait Models

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. I. Waller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Michael I. Waller

University of Wisconsin-Milwaukee

Paper presented at the Annual Convention
of the American Educational Research Association,
Boston, Massachusetts, April 1980

TM 800 664

Abstract

An objective method based on the likelihood ratio procedure is presented for use in selecting a measurement model from among the RASCH, 2-parameter and 3-parameter logistic latent trait models. The procedure may be applied in a straightforward manner to aid in choosing between the 2-parameter and the Rasch models. When choosing between the 3- and 2-parameter models, one must use the values of the parameters which maximize the likelihood on the restricted range $0 \leq \gamma_j \leq 1/A_j$; where γ_j is the guessing parameter and A_j is the number of alternatives. Also, given the inherent problem with the 3-parameter model of possible non-unique maximum likelihood parameter estimates and viable alternative procedures for handling random guessing, the recommendation is made that the 3-parameter model should be held in abeyance as a viable measurement model at this time.

An Objective Procedure for Comparing Logistic Latent Trait Models

Recently interest has developed in making comparisons among the three logistic latent trait models (Dinero & Haertel, 1977; Hutton, 1979; Reckase, 1978). While a number of issues must be considered when selecting a model for use with any particular measuring instrument, some of which are discussed below, it would be useful to have an objective technique available to aid in making the choice of model. This paper presents such an objective procedure based on the well-known likelihood ratio criterion (Wilks, 1962).

Briefly, the general likelihood ratio procedure may be described as follows: one estimates the full set of, say, r parameters of a model and calculates a likelihood ratio $\chi^2(r)$ goodness of fit statistic. Then one estimates only a subset of these parameters of size s using assumed values for the remaining p = r-s parameters and calculates a likelihood ratio $\chi^2(s)$. Finally, one calculates the difference between the two, $\chi^2(p) = (\chi^2(r) - \chi^2(s))$. This statistic, $\chi^2(p)$, can then be used to test the importance of estimating the p remaining parameters.

The Models

For a person with ability θ_i the 3-parameter model, the most general of the logistic latent trait models, may be written as in equation 1 (the superscript denotes the number of parameters in the model).

$$(1) \qquad P^3_{ij} = P_r \text{ (Person i correct on item j } \mid \theta_i)$$

$$= \gamma_j + (1-\gamma_j) \frac{e^{\beta_j - \alpha_j\theta_i}}{1 + e^{\beta_j - \alpha_j\theta_i}} \quad ;$$

where: γ_j is the guessing parameter or lower asymptote for item j,

β_j is the item difficulty of item j,

α_j is the item discrimination of item j.

θ_i is the ability of person i.

The next most general model, the 2-parameter model, assumes no guessing and the γ_j parameter for every item is set equal to zero. This model may be written as in equation 2.

$$(2) \qquad P^2_{ij} = \frac{e^{\beta_j - \alpha_j\theta_i}}{1 + e^{\beta_j - \alpha_j\theta_i}}$$

The Rasch or 1-parameter model further assumes that the items do not vary in discrimination and so the α_j parameter for every item is set equal to 1, yeilding the model as in equation 3.

$$(3) \qquad P^1_{ij} = \frac{e^{\beta_j - \theta_i}}{1 + e^{\beta_j - \theta_i}}$$

The value of the discrimination parameter in the Rasch model of equation 3 is slightly deceptive in that the discriminations are not necessarily one; rather, the discriminations can more accurately be said to be equal to some unknown constant value, α. The value of α is taken up as a scaling factor in the distribution of the underlving abilities, say θ^*_i, as shown in equation 4.

$$(4) \qquad P^1_{ij} = \frac{e^{\beta_j - \alpha\theta^*_i}}{1 + e^{\beta_j - \alpha\theta^*_i}} = \frac{e^{\beta_j - \theta_i}}{1 + e^{\beta_j - \theta_i}}$$

where: $\theta_i = \alpha\theta^*_i$

Estimation

Estimation of the parameters by maximum likelihood is accomplished in these models in a cyclic fashion. Each cycle is composed of two steps: estimation of the item parameters given the previous cycle's estimates of abilities; followed by estimation of the abilities given the previous cycle's estimates of the item parameters. These two-step cycles are continued until stabile item parameters are reached (see for example, Kolakowski & Bock, 1973; Wood, Wingersky & Lord, 1976; Wright & Mead, 1978).

In this paper we use the method developed by Bock and employed in the LOGOG computer program. In this procedure we estimate the item parameters under the assumption that the abilities follow a previously specified distribution; here we use the normal distribution with a mean and variance of zero and one respectively. This is accomplished at each cycle by taking the current estimates of abilities, ranking them, and distributing them into t groups or fractiles in such a way that the number of subjects, N_i, across the $i = 1, \ldots, t$ fractiles reflect the normal distribution. We then assume that within each fractile the subjects are sufficiently homogeneous so that we can proceed as though we have N_i independent observations all at the same ability level, θ_i, some middle value in fractile i. Formally, we proceed as follows:

Let $i = 1, 2, \ldots, t$ groups or fractiles whose subjects are sufficiently homogeneous as to be characterized by θ_i;

N_i = number of subjects in fractile i (determined by the assumed distribution);

r_{ij} = number of subjects in group i who respond to item j correctly;

$\underline{r}_j = \{r_{ij}\}$ = vector of item responses to item j across the t fractiles;

γ_j = the required guessing parameter;

β_j = the required difficulty parameter'

α_j = the required slope parameter

Then for a given item j the likelihood function is

$$(5)\quad L_j(\beta_j,\alpha_j,\gamma_j,|\underline{r}_j) = \prod_{i=1}^{t} \frac{N_i!}{r_{ij}!\ (N_i - r_{ij})!} P_{ij}^{\ r_{ij}}\ (1-P_{ij})^{(N_i-r_{ij})}$$

where P_{ij} = one of the three models specified above.

Therefore, the log likelihood for item j is given by

$$(6)\quad \ell_j = \log\left[\overline{L_j(\beta_j,\alpha_j,\gamma_j\ |\underline{r})}\right] = C + \sum_i^t \left[\overline{r_{ij}\log P_{ij} + (N_i-r_{ij})\log\ (1-P_{ij})}\right]$$

where $C = \log\ (N_i!/r_{ij}!(N_i-r_{ij})!)$

Given the matrices of first and second derivatives, we may apply Newton-Raphson iteration to the k^{th} stage estimate of the parameters to yield the $(k+1)^{st}$ stage estimates:

$$(7)\quad \begin{vmatrix}\gamma_j\\ \beta_j\\ \alpha_j\end{vmatrix}_{k+1} = \begin{vmatrix}\gamma_j\\ \beta_j\\ \alpha_j\end{vmatrix}_{k} + \begin{vmatrix}\ell_{\gamma_j\gamma_j} & \ell_{\gamma_j\beta_j} & \ell_{\gamma_j\alpha_j}\\ \ell_{\beta_j\gamma_j} & \ell_{\beta_j\beta_j} & \ell_{\beta_j\alpha_j}\\ \ell_{\alpha_j\gamma_j} & \ell_{\alpha_j\beta_j} & \ell_{\alpha_j\alpha_j}\end{vmatrix}_{k}^{-1} \begin{vmatrix}\ell_{\gamma_j}\\ \ell_{\beta_j}\\ \ell_{\alpha_j}\end{vmatrix}_{k}$$

(By eliminating the appropriate row(s) and column(s), the last equation can be modified for estimation in the 2 and 1-parameter models).

Procedure

For each item j a likelihood ratio χ_j^2 (m) goodness of fit statistic is calculated using the t fractiles as cells, and based on the log likelihood

of equations 5 and 6; that is,

(8) $\chi^2(m) = 2 \log (L_j) = 2 \ell_j$.

The degrees of freedom for this test statistic are t - m - 2; m equals the number of parameters estimated for the item and 2 additional degrees of freedom are lost since the mean and variance of the latent continuum are arbitraty and are set equal to zero and one, respectively. For a test of fit of the total measuring instrument, we may sum over the n item χ^2_j's yielding

(9) $$\chi^2_T(m) = \sum_j^n \chi^2_j (m)$$

which, for an m-parameter model, will be distributed asymptotically as a χ^2 on n(t-m-2) d.f..

We apply the general procedure to the problem of choosing among the 3 logistic latent trait models in the following manner. We may begin by assuming the 3-parameter model, estimating the parameters for all n items by maximum likelihood, and calculating $\chi^2_{T(3)}$. Given an adequate fit, we then perform the same calculations for the 2-parameter model yielding $\chi^2_{T(2)}$. At this point $(\chi^2_{T(3)} - \chi^2_{T(2)})$ is asymptotically distributed as a χ^2 on n d.f., and is a test of the null hypothesis.

$$H_o : \gamma_j = 0, \text{ all } j = 1, \ldots, n ;$$

against the alternative

$$H_1 : \gamma_j \neq 0, \text{ for at least one } j.$$

If we fail to reject H_o we may proceed to test that the Rasch model is adequate by performing the same calculations for this most restricted model yielding $\chi^2_{T(1)}$. Then, $(\chi^2_{T(2)} - \chi^2_{T(1)})$ is a test of $H_o : \alpha_j = \alpha$, all j=1,...,n ; against $H_1 : \alpha_j \neq \alpha$ for at least one j, α equal to some constant value.

There exist situations, e. g., affective measurement, where the guessing parameters can be assumed to be zero a priori. In such situations the focus becomes one of testing whether varying discriminations need be fit or whether the Rasch Model is adequate. To accomplish this test, we begin by assuming the 2-parameter model, estimating all the parameters and calculating $\chi^2_{T(2)}$. Given an adequate fit, we then perform the same calculations under the Rasch Model yielding $\chi^2_{T(1)}$. Now $(\chi^2_{T(2)} - \chi^2_{T(1)})$ is a test of

$H_o : \alpha_j = \alpha$, $j = 1,...n$; i. e., the Rasch is adequate against the alternative:

$H_1 : \alpha_j \neq \alpha$, for at least one j; i. e., the two parameter logistic is necessary to appropriately parameterize the measuring instrument.

A Monte Carlo Example

Two data sets were generated, each with N = 480 simulated subjects sampled from a normal distribution of abilities responding to n = 45 items. Data set R was generated simulating the Rasch Model with difficulties ranging from -4.4 to +4.4 in steps of .2 and with constant discriminations. Data set T was generated simulating the Two-Parameter Logistic Model with the same difficulties as data set R and with discriminations taken from the estimated discriminations of 45 items of an existing instrument; discriminations ranged from 0.8 to 2.7; mean = 1.56; standard devistion = .52.

Since guessing was assumed to be absent from the data in this example, we need only examine the fit of the 2-parameter and Rasch Models. Analyses producing maximum likelihood estimates of the parameters were performed using the LOGOG computer program for both data sets assuming each model; and likelihood ratio χ^2s were calculated in each of these four analyses. The results are presented in Table 1.

Insert Table 1 about here

From the first two columns of the table we see that, as expected, the two-parameter model produces an adequate fit to both sets of data; and that the one-parameter Rasch Model recovers Rasch data, but not two-parameter data. The third column contains the test for an adequate fit of the Rasch Model given that the two-parameter model fits the data: $(\chi^2_{T(2)} - \chi^2_{T(1)})$. Again as we would expect, this test correctly indicates that the Rasch Model analysis is appropriate for data which we know to be Rasch data and also correctly indicates significant lack of fit to the data which we know requires fitting two parameters for an adequate fit.

Discussion

In the example above, the analysis was straightforward. However, with data collected in situ, it is often the case that one would not obtain an adequate fit with the 2-parameter model in the first analysis; enough individual items might be found to fail to fit the multiparameter model so as to produce a significant lack of fit for the test as a whole, i. e., $\chi^2_{T(2)}$ would be found to be significant. In such a situation, standard item analysis techniques should be applied until a subset of items are found which satisfy the 2-parameter model; at that point, the Rasch Model would be applied to this subset of items and the test $(\chi^2_{T(2)} - \chi^2_{T(1)})$ performed to determine if the Rasch can appropriately be used.

The likelihood ratio procedure described in this paper is, of course, independent of the specific estimation procedure or computer program used. The computer program used for estimating the parameters of the model need not

calculate likelihood ratio chi squares; and, one can use different programs to estimate the parameters under the different models. Regardless of how the parameters have been estimated, one need only rank the subjects by their estimated abilities under each of the models under consideration, distribute them for each model into the same number of groups, and then calculate the log likelihoods as in equation (6) above for each item under each model. One then simply accumulates the item chi squares into $\chi^2_{T(m)}$ for each m-parameter model under consideration and performs the likelihood ratio tests described above using equations (8) and (9).

Of course, the decision concerning which model to use for a particular instrument should take into account other considerations as well. As is well-known, the raw score is a sufficient statistic for estimation of ability under the Rasch Model only (Birnbaum, 1968); and this characteristic implies a number of advantages which accrue to the Rasch. The most important advantage, perhaps, is that if a test developer or manufacturer obtains a set of items which fit the Rasch for a particular measuring instrument, future users of the instrument need not possess the resources of a computer in order to obtain estimated abilities as is required with the multiparameter logistic models. Instead, future users of such an instrument simply count the number correct and refer that number to a table developed by the test manufacturer containing the corresponding estimated abilities. These abilities lie on the well-defined metric inherent in every latent trait model.

This might suggest that when a test developer is in a situation where he has a pool of dichotomously scored items large enough to enable him to select a sufficient number of items all possessing a common discrimination, he may wish to use only the Rasch Model. On the other hand, sets of items all with a common discrimination are not readily available; indeed, most standardized

instruments available today contain items with varying discriminations (Lord & Novick, 1968). Consequently, if for any reason one is in a situation where one cannot readily discard items, one can take the set of items one has and employ the likelihood ratio test described in this paper to determine if the Rasch Model may appropriately be used with that set of items.

While the theory presented in this paper is straightforward for testing the adequacy of the Rasch Model given that the 2-parameter model fits the data, the test of the adequacy of the 2-parameter model given that the 3-parameter model fits requires additional comment. The third parameter, γ_j is introduced to deal with the problem of random guessing; and consequently is assumed to lie between 0 and $1/A_j$, where A_j equals the number of alternatives to item j. Mathematically, however, there is no such restriction on the γ_j parameter. In fact, maximum likelihood estimates of the guessing parameter are often negative and therefore seldom used. Rather, estimates of γ_j are obtained using other procedures; often, a graphic procedure is used in which one plots the proportion correct in each ability group or fractile and then graphically estimates the lower asymptote of the resulting logistic ogive. While restricting the range of the γ_j parameter does not prohibit one from using the statistical test described in this paper, in order to use this procedure one must use as estimates of all the parameters those estimates which maximize the likelihood on this restricted range of γ_j : $0 \leq \gamma_j \leq 1/A_j$.

Estimates of all the parameters under the restriction that the γ_j lies between 0 and $1/A_j$ would not be difficult to obtain provided we knew or could assume that the likelihood surface generated by the 3-parameter model is everywhere concave down. However no such assumption concerning the likelihood

surface may be made. Samejima (1973) has shown that with respect to estimation of abilities there exist individual response vectors for which the maximum likelihood estimate of θ_i given the item parameters is not unique; i. e., the maximum likelihood surface is not everywhere concave down for estimation of abilities from some response vectors. For estimation of the item parameters given the ability parameters the likelihood surface might be assumed to be similar, though not equivalent, to the shape of the likelihood for estimation of abilities. Therefore in the absence of mathematical evidence to the contrary, one can not generally assume a likelihood surface which is everywhere concave down.

This situation implies that in order to test the appropriateness of the 2- parameter model given the 3-parameter model fits the data, one must choose several values of γ_j between 0 and $1/A_j$ and then obtain maximum likelihood estimates of β_j and α_j and the value of the likelihood at each of these points in the parameter space.* One then chooses as the item parameters those values of α_j, β_j, and γ_j which produced the maximum value on the likelihood surface, and proceeds with the test as presented above.

A second implication which may be drawn from the general lack of a unique local maximum on the likelihood surface is that one might seriously question recommending the general use of 3-parameter model. The third parameter, γ_j is introduced primarily as a method for handling random guessing on multiple choice tests. There exist alternative procedures for handling this

---

* The next version of the LOGOG computer program will allow such estimation.

situation (Waller, 1974, 1976; Wright, 1977) which do not possess this limitation, and which are beginning to gain currency (Slinde & Linn, 1979, Smith & Mitchell, 1980). The lack of a unique local maximum would appear to be a severe limitation. Consider briefly the effect of this lack simply with respect to estimation of abilities. In addition to the lack of a unique local maximum likelihood estimate, Samejima's (1973) results indicate that for a single response vector, one could obtain two (or more) widely differing ability estimates simply by starting the iterative procedure with different initial estimates. Consequently, in order to obtain the one value of θ_i which produces the maximum value on the likelihood surface, one should examine the entire range of θ_i from $-\infty$ to $+\infty$ finding every local maximum AND the value of the likelihood at each such maximum. One could then choose as the ability estimate that value of θ_i which produces the largest value on the likelihood surface. To the author's knowledge, none of the computer programs presently available for estimation of the ability parameter under the 3-parameter model search in any way for the absolute maximum likelihood estimate of θ_i. In this light it would seem to be propitious to recommend that the 3-parameter model be held in abeyance as a viable latent trait model at least until such time as the problem of non-unique parameter estimates is resolved.

Summary

An objective method based on the likelihood ratio procedure is presented for use in selecting a measurement model from among the Rasch, 2-parameter, and 3-parameter logistic latent trait models. The procedure may be applied in

a straightforward manner to aid in choosing between the 2-parameter and the Rasch models. When choosing between the 3- and 2-parameter models, one must use the values of the parameters which maximize the likelihood on the restricted range: $0 \leq \gamma_j \leq 1/A_j$. Also, given the inherent problem with the 3-parameter model of possibly non-unique maximum likelihood parameter estimates and viable alternative procedures for handling random guessing, the recommendation is made that the 3-parameter model should be held in abeyance as a viable measurement model at this time.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical Theories of Mental Test Scores. Reading, MA: Addison Wesley Publishing Co., 1968.

Dinero, T. E., & Haertel, E. Applicability of the Rasch Model with varying items discriminations. Applied Psychological Measurement, 1977, 4, 581-592.

Hutton, L. An empirical comparison of the goodness of fit of three latent trait models to real test data. Paper presented at the Annual Convention of the American Educational Research Association, 1979.

Kolakowski, D., & Bock, R. D. LOGOG: Maximum likelihood item analysis and test scoring - Logistic model for multiple item responses. Chicago, IL: National Educational Resources, Inc., 1973.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison Weslel Publishing Co., 1968.

Reckase, M. D. A comparison of the one and three parameter logistic models for item calibration. Paper presented at the Annual Convention of the American Educational Research Association, 1978.

Samejima, F. A comment on Birnbaum's three-parameter logistic model in the latent trait theory. Psychometrika, 1973, 38(2), 221-234.

Slinde, J. A., & Linn, R. L. The Rasch Model, objective measurement, equating and robustness. Applied Psychological Measurement, 1979, 3(4), 437-52.

Smith, R. M., & Mitchell, V. P. An analysis of Rasch person fit statistics in college placement tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, April 1980.

Waller, M. I. Estimating parameters of the Rasch Model: Removing the effects of random guessing (ETS RB 76-8). Princeton, NJ: Educational Testing Service, 1976.

Waller, M. I. Removing the effects of random guessing from latent trait ability estimates (ETS RB 74-32). Princeton, NJ: Educational Testing Services.

Wilks. S. S. Mathematical Statistics. Tokyo, Japan: John Wiley & Sons, Inc., 1967.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research memorandum 78-6, Princeton, NJ: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14(2), 97-116.

Wright, B. D., & Mead, R. J. BICAL: Calibrating items and scales with the Rasch Model. Reserach Memorandum Number 23A, University of Chicago.

Table 1

Likelihood Goodness of Fit of 's

DATA MODEL		Estimation Model: Rasch χ^2	d.f.	Prob.	Estimation Model: 2-Parameter χ^2	d.f.	Prob.	Test of Rasch Fit Given 2-Parameter Fits $(\chi^2_{T(2)} - \chi^2_{T(1)})$ χ^2	d.f.	Prob.
Rasch		356.86	315	NS	298.96	270	NS	57.90	45	NS
2-Parameter		695.23	315	<<.00001	235.58	270	NS	459.65	45	<<.00001